DOCUMENT RESUME

ED 267 115 TM 860 206

AUTHOR Goodwin, Coleen R.
TITLE The Pitfalls of Testing in the Quest of Excellence.
PUB DATE Apr 85
NOTE 8p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Criterion Referenced Tests; Educational Assessment; *Educational Quality; Elementary Secondary Education; Measurement Techniques; Norm Referenced Tests; School Effectiveness; *Scores; *Testing; *Testing Problems; *Test Interpretation

ABSTRACT

This document presents five scenarios, each referring to a means of assessing learning. These examples are then used to discuss pitfalls faced when tests are used to assess learning, and the conclusion is reached that testing should play a supportive role only, not a major role in the quest for excellence. Effective schools research shows that testing is only a part of the assessment of student progress; moreover, more than one type of testing should be used. Criterion referenced tests are a better vehicle for assessing student progress in any given area, as compared with norm referenced tests (NRT). When NRTs are used, instead of looking at one mean score, it is necessary to look at test data in a variety of ways: (1) by parent (particularly mother's) education level; (2) by socio-economic status group; (3) by race; and (4) by aptitude. In the quest for academic excellence, many districts are now looking at a number of factors which are indicators of excellence rather than merely test scores. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

FD267115

The Pitfalls of Testing in the Quest of Excellence

Coleen R. Goodwin, Ed.D.
Supervisor of Research Services
Phoenix Union High School District
Phoenix, AZ 85017

A Symposium Presented at the Annual Meeting of the
National Council of Measurement in Education in Chicago, Illinois
April, 1985

TM 860 206

NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C. R. Goodwin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Scenario 1:

Study Seeks to Identify Exemplary Districts in Ohio

"More than 40 suburban school districts in Ohio - members of the Council for Academic excellence - are working with the Educational Records Bureau to identify districts where students score better on achievement tests than predicted by aptitude tests.

Based on Tests developed by ETS, the aptitude/achievement analysis will identify exemplary districts where achievement test scores exceed aptitude test predictions by the greatest extent.

Peter Vitella, ETS director of Educational Records Bureau programs, says the mean differences between actual and predicted achievement scores will be computed for grades 4, 7, and 10 in each of the participating districts.

'Those with the highest mean differences will be identified as the exemplary districts, and will share information about curriculum, instructional materials, media use, and teacher in-service programs,' Vitella says.

'This information constitutes a valuable tool in helping administrators and teachers identify exemplary achievement in various school districts,' he continues. 'For example, if fourth graders in a particular district are scoring consistently and significantly higher in reading comprehension than their verbal aptitude test scores predict, other districts will want to look for reasons for these positive results.....'

Vitella notes that tests provided through this program are administered by individual schools at their convenience. Special norms are provided for independent schools and suburban districts so school administrators can compare the performance of their students with that of students in similar schools."[1]

Scenario 2:

January 1984

Secretary Terrel Bell unveiled his now familiar wall chart that ranked states according to about a dozen factors. The variable that journalists seized and widely reported, however, was the deceptively simple, available, and familiar one of college entrance exam scores on the Scholastic Aptitude and American College Tests (SAT and ACT).[2]

Scenario 3:

"After several years of being stuck in task force reports and committee discussions, the words 'education indicators' are suddenly very much on the minds of education policymakers. The search for a few key touchstones that can be used to measure educational progress has come to the forefront of edcuational issues only within the last year, but the effects of indicators soon to be chosen may permanently alter the character of educational assessment in this country.

The Council of Chief State School Officers (CCSSO), a collective of state-level superintendents, recently came to a surprising decision to support the development of a system of indicators.

After lively debate and despite some chiefs' grave reservations, at their annual November meeting in Wilmington, Del., the council broke with its tradition of eschewing nationally standardized policy. The chiefs approved a plan committing themselves to development of national indicators that can be used by the public for state-by-state school comparisons and can be used by states to measure their educational progress. Test scores may be included as one indicator, suggesting the eventual possibility of a national test.

A list of proposed indicators is to be presented to the chiefs by the project staff at their next annual November meeting.

Stephen Kaagan, the Vermont commissioner who chaired the committee presenting the plan, said that like it or not, standardized testing is now used to evaluate education quality. What is needed, he said, is a broader look. He pointed out that the analysis of student writing is more sophisticated than it has ever been and speculated about using the voting behavior of 18- to 24-year-olds as a measure of how well schools foster citizenship.....

The move toward national measurement has now shifted from whether to proceed to how it will be done. NCES has reconvened its task force and is testing reaction to a long list of possible indicators which the task force hopes to pare down to 20 or so items."[3]

Scenario 4:

"Massachusetts' efforts to develop a practical and reliable method of judging students' speaking skills in a large-scale assessment have given impetus and encouragement to many local and state assessment programs. But for Massachusetts, refinements in the testing are still in process.

How it Worked

To recap the Massachusetts program, project staff - with the assistance of speech communication experts - devised a two-stage speaking assessment for a 1980 trial test of students' speaking competencies statewide. In the assessment, teachers first rated the general speaking abilities of the students in class, based on observations. Then trained raters conducted a one-on-one structured assessment for students whose initial rating was close to minimum passing standards.

The classroom ratings, which were the initial basis of the assessment, called for two teachers in different subject areas to independently rate each student's general speaking skills on a four-point scale. The raters considered delivery, organization, content and language. Then, for students with borderline competencies, a third rater judged the student's ability to carry out four specific speaking tasks: describing a person or a thing, explaining an activity in step-by-step sequence, requesting help in an emergency, and presenting a persuasive argument.

Reliability: Some Surprises

A field test in Spring 1980, with 12th graders randomly selected from 49 schools, provided promising results. Yet staff remained uncertain whether the procedures provided reliable information. Would individual student scores be the same if

students were assessed a second time? Or by different raters?

A systematic check of the procedures yielded some surprises: The classroom ratings were definitely unreliable, so much so that they could not be used as the basis of the assessment. On the other hand, results from the one-on-one ratings, based on structured exercises, were substantially more reliable. Moreover, the one-on-one ratings and classroom observations were not highly correlated. These findings encouraged staff to rethink the whole structure of the two-stage assessment. Although some argued that the teacher observations were a 'more valid, practical and direct method of assessing speaking,' data from the study suggested otherwise. In fact, the teacher observations 'were so unreliable that they were virtually useless.'"[4]

Scenario 5:

March 1985

All the teachers in Arkansas are given a test of their basic skills.

Question: What do these five scenarios have in common? Each refers to a means of assessing learning.

Question: What pitfalls do we face as tests are used to assess learning? Let's look at the pitfalls in the previous 5 scenarios.

Scenario 1: Pitfalls

If a district tries to emulate the ETS study by selecting tests already on the market, (i.e., not using the ETS specially prepared tests), inaccurate assumptions may be made. Tests on the market vary widely when administered to the same population. For example, this fall an equating study was conducted in Phoenix between two tests. The results were as follows:

Student achievement on (Test B) and (Test A) was compared in the areas of reading, language, and mathematics. The equating study showed that for reading in Grade 11, a (Test A) grade equivalent (GE) score of 11.7 (eleventh grade, seventh month) was approximately equal to a GE score of 9.8 (ninth grade, eighth month) on (Test B). An eleventh-grade student, who ranked at the 50th percentile for reading on (Test A) when compared to all students in the nation, only ranked at the 37th percentile on (Test B) when compared nationally. English and mathematics comparisons showed similar differences between (Test B) and (Test A) test scores.

These results show that (Test B) norms are tougher than (Test A) norms. From these equating study results, it is predicted that Arizona students will earn lower GE and National Percentile Rank (NPR) scores on (Test B) than they have on (Test A) in the past.

The reason given by the Arizona Department of Education was: "that these tests were normed on different students in different years. (Test A) norms were set in 1977; (Test B) norms were set in 1982. School achievement showed a noticeable increase during this five-year period and students had to perform better on 1982 tests to maintain GE and NPR scores equal to those of 1977."[5]

Scenario 2: Pitfalls — Tests are developed for specific purposes. Use of the ACT and SAT to measure knowledge acquired in high school is not a true measure of student learning. The two tests were developed as college aptitude tests and were normed on a specific population: college bound students. Only 72 percent of the 18-year-old population graduate from high school[6] and approximately 54 percent of those ever enter college.[7]

Scenario 3: Pitfalls — People do not realize the logistical nightmares of having a national exam. Consider the bureaucracy that it would create to administer the program. Consider the test security problems. Consider how "tricky" some test items are. Consider what vast cultural differences we have in the "melting pot" called America. Consider the vast ability differences of people. Consider how tests are normed. There will always be 50 percent below what is determined to be "grade level" if a norm-referenced test is used. Consider the consequences of such a report if one valid instrument could be prepared.

Scenario 4: Pitfalls — This scenario shows the problems faced when subjective judgment of teachers comes into the picture in evaluating student performance. Each teacher has a different educational background and each teacher has a different "value system" when it comes to judging quality of student performance. For years English teachers have struggled with this problem in grading student writing. It is very difficult to have consistency in teacher measurement of learning when there is no consistency in the training of teachers.

Scenario 5: Pitfalls — The logistical premise of the Arkansas teacher testing seems to be as follows:

Major premise: All teachers who pass the Arkansas test are good teachers.
Minor premise: Miss Jones passed the Arkansas test.

Conclusion: Therefore Miss Jones is a good teacher.

That is not true. We have all had teachers who were veritable encyclopedias of knowledge who could not teach students. They lacked the ability of transmitting their wealth of knowledge in understandable terms to others.

HOW SHOULD TESTING BE USED IN THE QUEST FOR EXCELLENCE?

Then what role should testing play in the quest for excellence? It should play a supporting role only, not a major role.

As the effective schools research has shown, there are many factors and elements that make up effective schools. Testing is only a part of the assessment of student progress. However not just one type of testing should be used. School districts who have had a lot of experience with both NRT's and CRT's realize that the CRT's are a better vehicle for assessing student progress in any given area.

An NRT is fine for taking a "temperature reading" of where the district is compared to others nationally on one specific test. But in order to improve, the schools must look at their areas of weakness as well as strength. Recent NRT's try to give skill analysis, but they are very narrow in the scope of what they can cover. Well-prepared CRT's can do the job for the schools.

When NRT's are used, instead of looking at one mean score, we need to look at test data in a variety of ways:

by parent (particularly mother's) education level
by SES group,
by race,
by aptitude, and so forth.

Then the data needs proper explanation to the various publics.

The consequences of the selection of the wrong test to measure the wrong thing are far reaching. We have to face the fact that many people look at test scores as the "final judgment." The horror stories concerning misclassification and misdiagnosis using test scores are many. One problem large cities face is that when certain schools get a certain academic reputation, it is hard to remove the image from the public's mind. Real estate salesmen perpetuate the problem and as a result neighborhoods have changed rapidly.

Exemplary districts/schools are not necessarily those whose only "claim to fame" is high test scores. The fact might be that those test scores are not as high as they could be.

Exemplary districts/schools should be those that are literally "pulling themselves up by their boot straps."

Conclusion:

The problem of educating the public, including the media, as to the proper use and interpretation of test results is a never ending battle. In fact, it is a battle that needs to be better fought. Too many laws are being legislated wherein those who write the laws are unaware of the consequences of what they have mandated.

One hope we have is that in the quest for excellence, many districts are now looking at a number of factors which are indicators of excellence rather than merely test scores. As districts/schools start broadening their base of reporting on indicators, better true assessments of learning will be forthcoming.

Cited Sources

1. ETS Developments, Autumn 1984, p. 10

2. ASCD Update, Vol. 27, Number 1, January 1985, p. 7

3. ASCD Update, Vol. 27, Number 1, January 1985, p. 1, 7

4. CAPTRENDS, Vol. 9, Number 2, February 1984, pp. 3-4

5. Arizona Department of Education Memorandum, March 20, 1985

6. The Condition of Education 1984, National Center for Education Statistics, p. 13

7. The Condition of Education 1982, National Center for Education Statistics, p. 31